Metabarcoding and Metagenomics 6: 245–259
DOI 10.3897/mbmg.6.85794

Research Article

# Towards quantitative metabarcoding of eukaryotic plankton: an approach to improve 18S rRNA gene copy number bias

**Jon Lapeyra Martin**[1], **Ioulia Santi**[2], **Paraskevi Pitta**[3], **Uwe John**[4], **Nathalie Gypens**[1]

1 *Laboratoire d'Ecologie des Systèmes Aquatiques, Université Libre de Bruxelles, CP221, Boulevard du Triomphe, Brussels 1050, Belgium*
2 *Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), P.O. Box 2214, GR71003 Heraklion, Crete, Greece*
3 *Hellenic Centre for Marine Research (HCMR), Institute of Oceanography (IO), P.O. Box 2214, GR71003 Heraklion, Crete, Greece*
4 *Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Ammerländer Heerstraße 231, 26129 Oldenburg, Germany*

Corresponding author: Jon Lapeyra Martin (jon.lapeyra.martin@ulb.be)

**Academic editor:** Thorsten Stoeck | **Received** 25 April 2022 | **Accepted** 1 August 2022 | **Published** 15 August 2022

## Abstract

Plankton metabarcoding is increasingly implemented in marine ecosystem assessments and is more cost-efficient and less time-consuming than monitoring based on microscopy (morphological). 18S rRNA gene is the most widely used marker for groups' and species' detection and classification within marine eukaryotic microorganisms. These datasets have commonly relied on the acquisition of organismal abundances directly from the number of DNA sequences (i.e. reads). Besides the inherent technical biases in metabarcoding, the largely varying 18S rRNA gene copy numbers (GCN) among marine protists (ranging from tens to thousands) is one of the most important biological biases for species quantification. In this work, we present a gene copy number correction factor (CF) for four marine planktonic groups: Bacillariophyta, Dinoflagellata, Ciliophora miscellaneous and flagellated cells. On the basis of the theoretical assumption that '1 read' is equivalent to '1 GCN', we used the GCN median values per plankton group to calculate the corrected cell number and biomass relative abundances. The species-specific absolute GCN per cell were obtained from various studies published in the literature. We contributed to the development of a species-specific 18S rRNA GCN database proposed by previous authors. To assess the efficiency of the correction factor we compared the metabarcoding, morphological and corrected relative abundances (in cell number and biomass) of 15 surface water samples collected in the Belgian Coastal Zone. Results showed that the application of the correction factor over metabarcoding results enables us to significantly improve the estimates of cell abundances for Dinoflagellata, Ciliophora and flagellated cells, but not for Bacillariophyta. This is likely to due to large biovolume plasticity in diatoms not corresponding to genome size and gene copy numbers. C-biomass relative abundance estimations directly from amplicon reads were only improved for Dinoflagellata and Ciliophora. The method is still facing biases related to the low number of species GCN assessed. Nevertheless, the increase of species in the GCN database may lead to the refinement of the proposed correction factor.

## Key Words

18S rRNA copy number, correction factor, plankton, protists, quantitative metabarcoding

## Introduction

During the last decades, unicellular eukaryotic plankton has been used as an indicator of ecosystem change due to its rapid response to environmental variations (e.g. Pawlowski et al. 2016). Monitoring programs have usually relied on microscopy, which is time-consuming and requires high taxonomic skills (Stern et al. 2018). While microscopy allows the identification based on morphology (Edler and Elbrächter 2010) as well as the enumeration and carbon biomass (C-biomass) estimation (Menden Deuer and Lessard 2000) of microorganisms (usual-

ly >20 µm), molecular approaches such as High Throughput Sequencing (HTS) have a high potential for detailed species monitoring (Ebenezer et al. 2012; de Vargas et al. 2015; Stern et al. 2018), capturing the entire size-range of the protistan community including nano- and picoplanktonic components (Elferink et al. 2017, 2020; Bruhn et al. 2021). Deoxyribonucleic acid (DNA) metabarcoding has proven to be a powerful and sensitive tool for large-scale biodiversity surveys, allowing comparison of studies rooted in taxonomy (Chain et al. 2016). Although considered as an attractive approach to assess protist diversity in nature (Medlin and Kooistra 2010; Santoferrara et al. 2020) it is subject to several distinct biases (technical and biological) that influence sequence read counts and estimated diversity (Thomas et al. 2016). These uncertainties limit to some degree its application for biomonitoring.

The small ribosomal subunit (SSU) 18S rRNA gene is the most widely used marker for the detection and classification of aquatic eukaryotic protists. The different gene regions such as V9 (de Vargas et al. 2015) or the more lately used V4 (Piredda et al. 2017; Armeli Minicante et al. 2019) offer conserved primer binding sites that are used to amplify broad taxonomic groups via polymerase chain reaction (PCR), providing some degree of taxonomic resolution. In order to understand the microbial diversity, species succession or dynamics, several ecological studies based on metabarcoding datasets (Massana et al. 2015; Armeli Minicante et al. 2019; Gran-Stadniczeñko et al. 2019; Käse et al. 2020; Bruhn et al. 2021; Lapeyra Martin et al. 2022) have commonly relied on the acquisition of organismal abundances directly from the number of DNA sequences (i.e. reads). These reads are later assigned to Operational Taxonomic Units (OTUs) or Amplicon Sequences Variants (ASVs), from which relative abundances of the community are ultimately calculated.

Quantification of marine protists based on ASV/OTU relative abundances has been largely discussed in the literature (Weisse et al. 2016; Vasselon et al. 2018; Santoferrara 2019; Käse et al. 2020). There exist several inherent technical issues occurring due to sample preservation (Mäki et al. 2017), DNA extraction (Van der Loos and Nijland 2021), primer choice and specificity (Elbrecht and Leese 2015; Lapeyra Martin et al. 2022; Latz et al. 2022) and ultimately PCR, which under- or overestimate different groups (Wintzingerode et al. 1997; Gonzalez et al. 2012; Latz et al. 2022). Besides the technical aspect, various authors agree that the major source of bias avoiding an accurate metabarcoding quantification is the 18S SSU rRNA gene copy number (GCN) variation within species, genera and plankton groups (Not et al. 2009; Mäki et al. 2017; Vasselon et al. 2018; Saad et al. 2020). Deviation in GCN between species or even individuals is well documented (Zhu et al. 2005; Gong et al. 2013) and can be substantial, affecting the proportion of reads found for each species present in complex environmental assemblages. This often leads to misinterpretation of relative abundances when comparing with proportions revealed by microscopic counts (Santi et al. 2021).

In the case of prokaryotes, phylogeny-based approaches have been applied to estimate 16S rRNA GCN and potentially correct this bias (Kembel et al. 2012; Angly et al. 2014). In the case of eukaryotes, a limited number of genomes have been sequenced and few species-specific 18S rRNA GCN have been assessed (Yarimizu et al. 2021), making the same corrective approach an arduous task. Moreover, it has been observed that 18S rRNA GCN, cellular biovolume and carbon content relationship strongly vary between different taxonomic groups and species (Lee et al. 2009; Galluzzi, Penna 2013; Mäki et al. 2017; Gong and Marchetti 2019). Particularly, differences have been already demonstrated for dinoflagellates (LaJeunesse et al. 2005; Galluzzi and Penna 2013; Toebe et al. 2013; Yarimizu et al. 2021), diatoms (Connolly et al. 2008; Godhe et al. 2008), ciliates (Gong et al. 2013) and other flagellates (Zhu et al. 2005; Read et al. 2013). Rigorous comparisons of morphological (microscopy) and metabarcoding methodologies exist in the literature for eukaryotic plankton in freshwater environments (Medinger et al. 2010; Groendahl et al. 2017), estuarine ones (Abad et al. 2016), the Mediterranean Sea (Piredda et al. 2017; Santi et al. 2021), the North Sea (Käse et al. 2020) or the Skagerrak basin (Gran-Stadniczeñko et al. 2019). These studies highlighted the incongruences of the results of both approaches (microscopy and metabarcoding), that might be occurring due to 18S rRNA GCN differences among taxa, as for example Alveolate (e.g. ciliate and dinoflagellate) sequences usually constitute the largest fraction of sequence reads.

Some approaches have been lately developed to assess and/or mitigate quantification bias in metabarcoding: making use of control material in fish (Thomas et al. 2016) or plants (Matesanz et al. 2019), reducing amplification bias in arthropods (Krehenwinkel et al. 2017) and zooplankton (Ershova et al. 2021), or use of biovolume for some protistan plankton taxa such as diatoms (Vasselon et al. 2018). However, since we still lack approaches that address the inherent issue of quantitative metabarcoding for the entire eukaryotic plankton community, the use of relative abundances directly obtained from the number of reads in marine protistan plankton assemblages collected from environmental samples is yet relatively common.

Therefore, assuming that the variation in relative abundances of different organisms and/or taxa can be partly attributed to the natural inherent differences in 18S rRNA GCN, the present study attempts to investigate the use of a GCN correction factor (GCN-CF) to mitigate the quantitative bias (cell number and biomass) in DNA metabarcoding approaches applied to marine protistan plankton surveys. We therefore compared the quantitative discrepancy between microscopic and molecular methods aiming to answer the following question: can we improve the estimations of cell and biomass relative abundances from metabarcoding reads if we consider the taxa-specific 18S rRNA GCN?

In order to answer this question (1) we compiled marine unicellular eukaryotic plankton data available in published literature of 18S rRNA GCN, (2) we created

a taxa-specific GCN-CF from the GCN database (GCN database) and (3) we applied the GCN-CF to the metabarcodes of 15 environmental DNA (eDNA) samples collected in the Belgian Coastal Zone (BCZ). Corrected metabarcodes were compared to their corresponding microscopical counts and biomass estimations in relative abundances.

# Materials and methods

We focused on four single cell marine eukaryotic plankton groups (protists) commonly determined and enumerated under the microscope (Edler and Elbrächter 2010; Manoylov 2014): diatoms, dinoflagellates, ciliates and various flagellated cells. Their corresponding taxonomic ranks according to the National Center for Biotechnology Information (NCBI) taxonomy are: Bacillariophyta (Stramenopiles), Dinoflagellata (Alveolata), Ciliophora (Alveolata). The fourth group used in this study, flagellated cells, is a non-taxonomic term that groups a wide range of microbial eukaryotic cells that have flagella and are not diatoms, dinoflagellates, ciliates and are usually <20 µm in size. In the present study, organisms classed within flagellated cells may correspond to species belonging to the following listed groups: Bigyra, Opalozoa, Cercozoa, Chlorophyta, Cryptophyta, Haptophyta, Hyphochytriomycota, Labyrinthulomycetes, Ochrophyta, Oomycota, Rhodophyta, Foraminifera, Radiolaria, and Apicomplexa.

## Gene copy number & correction factor

We built up a species-specific 18S rRNA GCN database for marine protists that contained species specific GCN per cell values, along with the estimated cellular biovolume and carbon content (C-content), which were acquired from several studies published in the literature and are listed in Suppl. material 1. The absolute GCN per cell in the original research articles were achieved through different methodological approaches such as qPCR (Zhu et al. 2005), real-time PCR (Godhe et al. 2008), single-celled qPCR (Gong et al. 2013), bioinformatics pipeline (Gong, Marchetti 2019) and single cell digital PCR (Yarimizu et al. 2021). We assumed that the different 18S rRNA GCNs listed in the GCN database (Suppl. material 1) are the absolute 18S rDNA gene copy numbers per cell and that each GCN is read once during one metabarcoding sequencing event, and therefore that '1 read count' is equivalent to '1 GCN'.

From the GCN database, we calculated the mean, median, and standard deviations GCN per cell for each of the four established groups: Bacillariophyta, Dinoflagellata, Ciliophora and flagellated cells. Median GCN cell$^{-1}$ values were used due to data limitation and presence of outliers. These values per group were used as key components for the correction factors and development of the mathematical equations.

The equation (i) estimates the corrected metabarcoding cell relative abundances (MTB_CFcell) for each defined plankton group (g) within a single sample. Metabarcoding number of reads (expressed in GCN) are divided by the corresponding plankton group CF (GCN cell$^{-1}$), and the following division by the total sum of the four groups estimates ultimately the proportional contribution to the community (%).

The equation (ii) estimates the cellular GCN:C-content ratios (CC ratio) for each plankton group (values reported in Table 1). Median cellular C-content values (pg C cell$^{-1}$) were calculated from the microscopy (MCP) dataset for each plankton group, in order to be consistent with cellular C-content of the species found in the area (BCZ). Finally, the equation (iii) estimates the corrected C-biomass relative abundances (MTB-CFbio).

$$g = (\text{Bacillariophyta, Dinoflagellata, Ciliophora, Flagellated cells})$$

$$MTB\_CFcell_g[\%] = \frac{\frac{MTB_g[GCN]}{CF_g[GCN.cell^{-1}]}}{\sum_g\left(\frac{MTB_g[GCN]}{CF_g[GCN.cell^{-1}]}\right)} \quad \text{(i)}$$

$$CC\ ratio_g[GCN.pg\ C^{-1}] = \frac{CF_g[GCN.cell^{-1}]}{C\ content_g[pg\ C.cell^{-1}]} \quad \text{(ii)}$$

$$MTB\_CFbio_g[\%] = \frac{\frac{MTB_g[GCN]}{CC\ ratio_g[GCN.pg\ C^{-1}]}}{\sum_g\left(\frac{MTB_g[GCN]}{CC\ ratio_g[GCN.pg\ C^{-1}]}\right)} \quad \text{(iii)}$$

The relative abundances of the uncorrected metabarcoding results and the corrected ones from equation (i) and (iii), MTB_CFcell and MTB_CFbio respectively, were directly compared to MCP. This comparison was performed for the fifteen environmental samples.

## Field sampling

Seawater samples were collected at 3 m depth using 4 L Niskin bottles connected to a CTD sensor (Sea-bird SBE25). The monitoring in the Belgian Coastal Zone (Lapeyra Martin et al. 2022) took place from March 2018 to June 2019 (see Suppl. material 4: Table S2) aboard the RV Simon Stevin (Vlaams Instituut voor de Zee) at the St. 330 (51°26.05'N, 02°48.50'E). Both in 2018 and 2019, throughout the spring-summer months, one extra monthly cruise was undertaken and samples were collected to closely follow the evolution of the phytoplanktonic blooms that occur during this time period in Belgian waters (Gypens et al. 2007).

## Metabarcoding

The DNA samples for the study of the protistan community were collected vacuum filtering 500–800 mL of water (from Niskin) through 0.22 µm polycarbonate filters (47 mm) and storing the samples immediately at -20 °C. Total DNA was extracted from filters using NucleoSpin Soil extraction Kit (Macherey-Nagel, Düren, Germany)

following manufacturer's protocol. For a maximum efficiency of the extraction from the filters, a sample lysis step was added using 10 mL cryotubes (using a high velocity bead beater for 10 min). Up to three filters were pooled and used for DNA extraction when no sufficient biomass was found on a single filter. Standard polymerase chain reactions (PCR) were performed to amplify the universal eukaryote SSU 18S rRNA gene. Primers TAReuk454F-WD1 (5′-CCAGCASCYGCGGTAATTCC-3′), TAReukREV3 (5′-ACTTTCGTTCTTGATYRA-3′) were used to target the V4 region of the 18S rRNA gene (Stoeck et al. 2010). PCR reactions performed had a total volume of 25 µL, containing 2.5 µL of microbial DNA (5 ng $\mu L^{-1}$), 5 µL of both amplicon forward and reverse primers (1 µM) and 12.5 µL of high-fidelity polymerase HotStart ReadyMix (Kapa Biosystems). Plates were sealed and the following PCR-program was run in a thermal cycler: initial denaturation at 95 °C for 3 min, followed by 25 cycles of 95 °C for 30 s, annealing at 55 °C for 30 s; extension at 72 °C for 30 s final extension at 72 °C for 5 min. All PCR products (480 bp, ~383 bp + 97 bases of primers) were verified on a 1.5% agarose gel. The following library preparation of 18S ribosomal RNA gene amplicons was performed: PCR clean-up 1, index PCR, PCR clean-up 2, library quantification, normalization and pooling following the 16S Metagenomic Sequencing Library Preparation guide (Illumina 2013). Library denaturing and sample loading to the Illumina MiSeq system was performed to perform a 300 bp paired-end sequencing using V2 chemistry.

The reads were denoised and merged with DADA2, v1.16. (Callahan et al. 2016) using default cut-off parameters and annotation reads were subsequently classified with assignTaxonomy, the DADA2 implementation of the naive Bayesian classifier method described in Wang et al. (2007), against the Protist Ribosomal Reference Database PR2 v4.12.0 (Guillou et al. 2013). As this study is focusing on the protists, all reads assigned to Metazoans were excluded from the processed ASV-tables. Preparation of the ASV-tables was done in R v4.0 (R Core Team 2018) and assigned ASV taxonomy was used as proxy to classify the number of reads into described groups. The number of final sequencing reads ranged between 23,299 and 52,256 per sample. Subsampling to equal read number was done with the function rarefy from the 'vegan' package v2.0-10 (Oksanen et al. 2013). The metabarcoding dataset used for this study is available in Suppl. material 2 and can be found in online repositories: at DDBJ/EMBL/GenBank under the accession KFLC00000000. The raw data corresponding to the raw fastq files can be found on the online repository ZENODO, with the accession number: 10.5281/zenodo.6827112.

## Microscopy, biovolume and biomass

Seawater samples from St. 330 (100 mL) were fixed with Lugol's solution (1% final concentration), stored in the dark until microscope analysis using Utermöhl-type sedimentation (Edler and Elbrächter 2010). Cell counting and taxonomic identification down to genus/species level were performed under inverted light microscopy (LM) for all samples. If cells could not be identified down to the species or genus level using LM, higher taxonomic levels were used.

The cellular C-biomass was calculated based on biometric factors determined for each species and cell density. Mean species-specific biometric values were obtained from published data (Olenina et al. 2006; Nohe et al. 2018) and converted to biovolumes. Calculation of biomass was performed using carbon to volume relationships for diatoms, dinoflagellates and other protist plankton found in Menden-Deuer and Lessard (2000). In the case of ciliates, cell sizes were converted into cell volumes using geometric formulae from Peuto-Moreau (1991). Biovolumes were converted using the conversion factor 0.14 pg C $\mu m^3$ according to Putt and Stoecker (1989). For *Phaeocytis globosa*, a characteristic bloom forming species of the BCZ, three different life stages were identified and distinguished: flagellate, free-living and colonial. Biovolumes and cellular carbon content were estimated following species-specific carbon content data (Rousseau et al. 1990).

For each sample, proportional cell and C-biomass abundances from the microscopy dataset were calculated (percentage of the total cells and C-biomass per defined plankton group) and directly compared to the corrected and non-corrected metabarcoding data (i.e., relative abundance of each taxonomic group, estimated as the percentage of sequencing reads assigned to a taxonomic group compared to the total reads per sample). The microscopy results dataset used for this study is available in Suppl. material 3.

## Statistical analysis

All statistical analyses, data processing and plotting were performed using R version 4.0.2 (R Core Team 2018) and ggplot2 (Wickham, 2011) was used for visualization. Kruskal-Wallis and Wilcoxon tests were used to check differences in 18S rRNA GCN and C-biomass among different taxonomic groups or methods. The significance level was set at $p < 0.05$. Log-log plot was used to reflect the relationship over many orders of magnitude between GCN and biomass of marine protists. To check for differences in the community composition produced by the CF and assess its effects, we applied and compared generalized linear modeling based on beta distribution using the beta regression package *betareg* v3.1-4 (Cribari-Neto and Zeileis 2010). Two beta regression analyses were performed to assess the efficiency of the CF to improve cell relative abundances. The first one included the proportions of microscopy cell abundances and metabarcoding reads. The second model, instead of metabarcoding reads percentages, included the corrected results; after the application of the correction factor. As our dataset also assumes the "zero" values (referring to 0% of relative

abundance), thus a useful transformation in practice was used for the models: $(y \cdot (n - 1) + 0.5)/n$ where $n$ is the sample size (Smithson and Verkuilen 2006). The percentage of relative abundances of the eukaryotic groups were the response variable (y) and were tested against the explanatory variables (x), methods used (i.e.: MCP, MTB, MTB-CFcells, MTB-CFbio), and interaction of the methodology with each of the examined taxonomic groups.

# Results

## GCN database, metabarcoding and microscopy

The 18S rRNA GCN database comprised a total of 65 species' absolute GCNs per cell, biovolumes and estimated biomasses (see Suppl. material 1, 3). Species were classified in four taxonomic groups: Bacillariophyta (Stramenopiles), Dinoflagellata (Alveolata), Ciliophora (Alveolata) and flagellated cells. GCN data collected and sorted by taxonomic group are visualized in Fig. 1 and values reported in Table 1. Ciliophora showed the highest GCN cell$^{-1}$ followed by Dinoflagellata and Bacillariophyta, and flagellated cells being the group with the lowest GCN per cell on average by several orders of magnitude. Indeed, GCN per taxum showed significant differences among all taxonomic groups (Kruskal-Wallis, $p < 0.01$).

Regarding the cellular carbon content, a significant positive correlation ($R^2 = 0.79$; $p < 0.001$) was found in the GCN database dataset between the gene copy numbers and the estimated C-biomass of the single cell eukaryotes. When the same analysis was performed taking into consideration the taxonomic groups established (Fig. 1B), all groups showed a significant positive relationship with Ciliophora being the least correlated one ($R^2 = 0.47$, $p = 0.014$) (See Suppl. material 1 for references).

The comparison of the data produced by the three datasets used in this study, sorted by taxonomic groups, is shown in Suppl. material 4: Table S1. The comparison of the number of species per taxum included in metabarcoding (n = 1884) outnumbered microscopy (n = 165) and GCN database contained the smallest number of species (n = 65). Suppl. material 4: Fig. S1 displays a comparison of cellular C-content estimates between microscopy and GCN database. No significant differences were found for Bacillariophyta cellular C-content between both datasets. In fact, Bacillariophyta median cellular C-biomass was found to be almost equal in microscopy dataset and GCN database (37.6 and 37.9 pg C.cell-1 respectively). Regarding flagellated cells, even though the median and mean C-contents per cell were higher in GCN database, there was no significant difference found (Wilcoxon, $p > 0.05$). Ciliophora showed the highest difference of all groups both in median and mean biomass values.

**Table 1.** Plankton group specific correction factors (expressed in gene copy number per cell) cellular gene copy number (GCN) and GCN:C-content ratios (CC ratio) obtained with the application of equation (ii) for the Belgian Coastal Zone microscopy sample set.

| Plankton group | Correction Factor (GCN cell$^{-1}$) | CC ratios (GCN:pg C) |
|---|---|---|
| Bacillariophyta (Stramenopiles) | 166 | 4.41 |
| Ciliophora (Alveolata) | 71710 | 3.26 |
| Dinoflagellata (Alveolata) | 4919 | 27.17 |
| Flagellated cells | 5.23 | 0.94 |

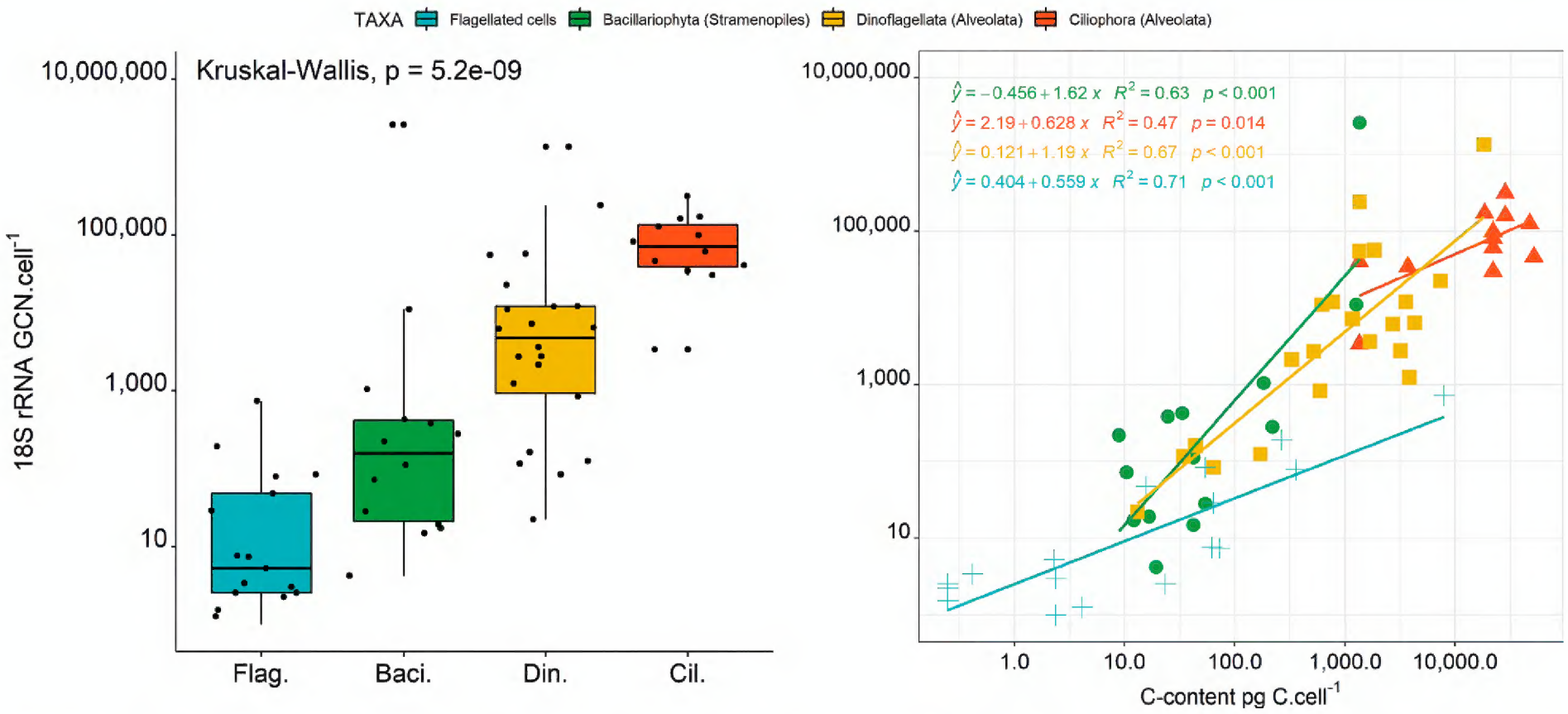


**Figure 1.** (**A**) 18S rRNA gene copy number per cell belonging to the major single celled eukaryotic plankton groups: Bacillariophyta (Stramenopiles), Dinoflagellata (Alveolata), Ciliophora (Alveolata) and flagellated cells. See Supplementary Table 1 for the median, mean and standard deviation values. (**B**) Log-log correlation between the number of 18S rRNA copies and cellular C-content (pgC cell$^{-1}$). The linear regressions were significant ($p < 0.05$) for all groups. Data were obtained through different methodological approaches: qPCR (Zhu et al. 2005), real-time PCR (Godhe et al. 2008), single-celled qPCR (Gong et al. 2013), bioinformatics pipeline (Gong, Marchetti 2019) and single cell digital PCR (Yarimizu et al. 2021).

Dinoflagellata C-biomass per cell showed significant differences among microscopy and GCN database (Wilcoxon, $p < 0.001$). In Table 1, cellular 18S rRNA GCN values are shown (displayed in Fig. 1A). The difference between the median and mean in GCN values was noted (emphasized specially in Bacillariophyta), where means were dominated by the outliers rather than the typical values. The median GCN values per taxum were used in this study as key components for the application of CF.

## Cell relative abundances

The cell relative abundances for the described plankton groups belonging to the 15 samples set from the BCZ are displayed in Fig. 2. Relative abundances refer to the evenness of distribution among defined groups in the community. Whereas the relative abundances in microscopy (cells) throughout all samples were dominated by the group flagellated cells, in metabarcoding proportions (reads) Dinoflagellata was the most abundant group (avg. = 51.4 ± 19.1%).

Bacillariophyta relative abundances from microscopy ranged between 1.4 – 61.9%, with an average (avg.) of 16.3 ± 14.7% throughout the total set (Fig. 2A). Corresponding metabarcoding proportions presented similar values with a mean 14.8 ± 7.8%, but with a maximum of 30.15% (Sample 15). Sample 3 showed the highest difference among microscopy and metabarcoding relative abundances (produced by blooming diatom *Minutocellus scriptus*, see Suppl. material 2). It was noted that the relative abundances between metabarcoding and microscopy results did not follow the same trend between samples. Dinoflagellata microscopy values were generally low ranging 0.08 – 2.8% (Fig. 2B), metabarcoding relative abundances were significantly higher varying between 77.2% (Sample 3) and 19.0% (Sample 9); reaching values higher than 50% in 9 out of 15 samples. The average of all samples for Dinoflagellata was the highest

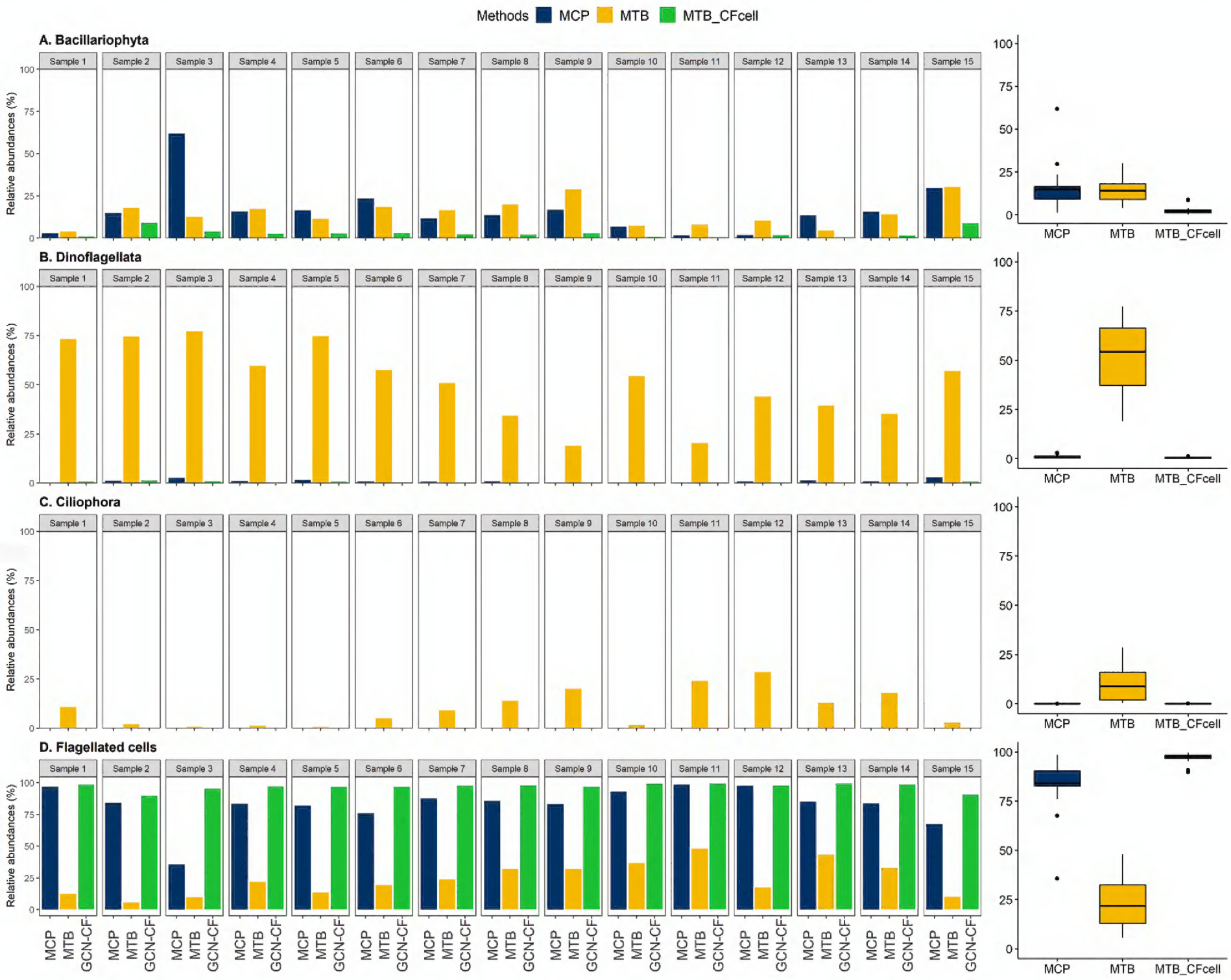


**Figure 2.** Comparison between relative abundances (%) of the cell counts from microscopy (MCP; blue), reads from DNA metabarcoding (MTB; yellow) and corrected read abundance percentages (MTB_CFcell; green) per sample. Right panels: box plots summarizing the comparison percentages of cell counts (MCP), metabarcoding reads abundance (MTB) and the corrected (MTB_CFcell) out of the 15 marine water samples per plankton group: (**A**) Bacillariophyta (Stramenopiles), (**C**) Dinoflagellata, (**B**) Ciliophora and (**D**) flagellated cells. The vertical line inside the box plots represents the median, the top and the bottom hinges correspond to the interquartile range, and the whiskers show the minimum and maximum non-outlier values. Note the scale difference between plankton groups.

among all groups and attained up to 51.4%. The case of the other Alveolates group studied, Ciliophora, was similar (Fig. 2C) with averaged microscopy values less than 1% and microscopy varying between 28.4% (Sample 12) and 0.53% (Sample 5). The group of flagellated cells was formed by organisms from different taxonomic groups generally < 20 μm. Despite the limitations for the adequate characterization of small organisms by the microscopy methodology, flagellated cells percentages (Fig. 2D) reached the highest values among the four groups ranging from 35.6% (Sample 1) to 98.4% (Sample 9) (avg. = 82.7 ± 15.4%). Microscopy values were higher than metabarcoding for all samples, metabarcoding values ranging between 5.6 – 47.9% (avg. = 23.8 ± 12.9%).

When the correction factor was implemented, the estimation of relative abundances decreased considerably (avg. = 2.7 ± 2.6%) for Bacillariophyta (Fig. 2A) in all samples. The application of the CF for Dinoflagellata, which strongly dominated the community structure in the metabarcoding dataset (Fig. 2B), was able to emulate more accurately the microscopy proportions decreasing the mean to ~ 0.4% (range = 0.04 – 1.2%). The relative abundance of metabarcoding reads attributed to Ciliophora was considerably higher than percentages obtained through MCP. Corrected values' average was found to be 0.03 ± 0.02%, getting much closer to the proportions found in MCP. When the CF was applied, the values for flagellated cells strongly approached the microscopy cell proportion values (avg. = 96.8 ± 2.8%).

Distributions of the two beta regression analyses, performed to assess the efficiency of the CF to improve cell relative abundances, were significant ($p < 0.01$) and detailed results are shown in Suppl. material 4: Table S3. According to the beta regression analysis containing the corrected data, GCN-CFcell resulted the best fit out of the two models with a higher $R^2$ and Phi precision parameter ($R^2 = 0.84$ to 0.96 and Phi ($\varphi$) = 14.96 to 33.23; displayed in Suppl. material 4: Fig. S2). The models significantly explained the variance in a different way for each phylum; this could be observed as well when examining the relative abundance estimates for each group in Fig. 2.

## C-biomass relative abundances

Equation (ii) was applied to calculate the cellular GCN:C-content ratios for the species present in the microscopy dataset of the BCZ and are reported in Table 1. The ratios varied from 0.94 for flagellated cells to 27.17

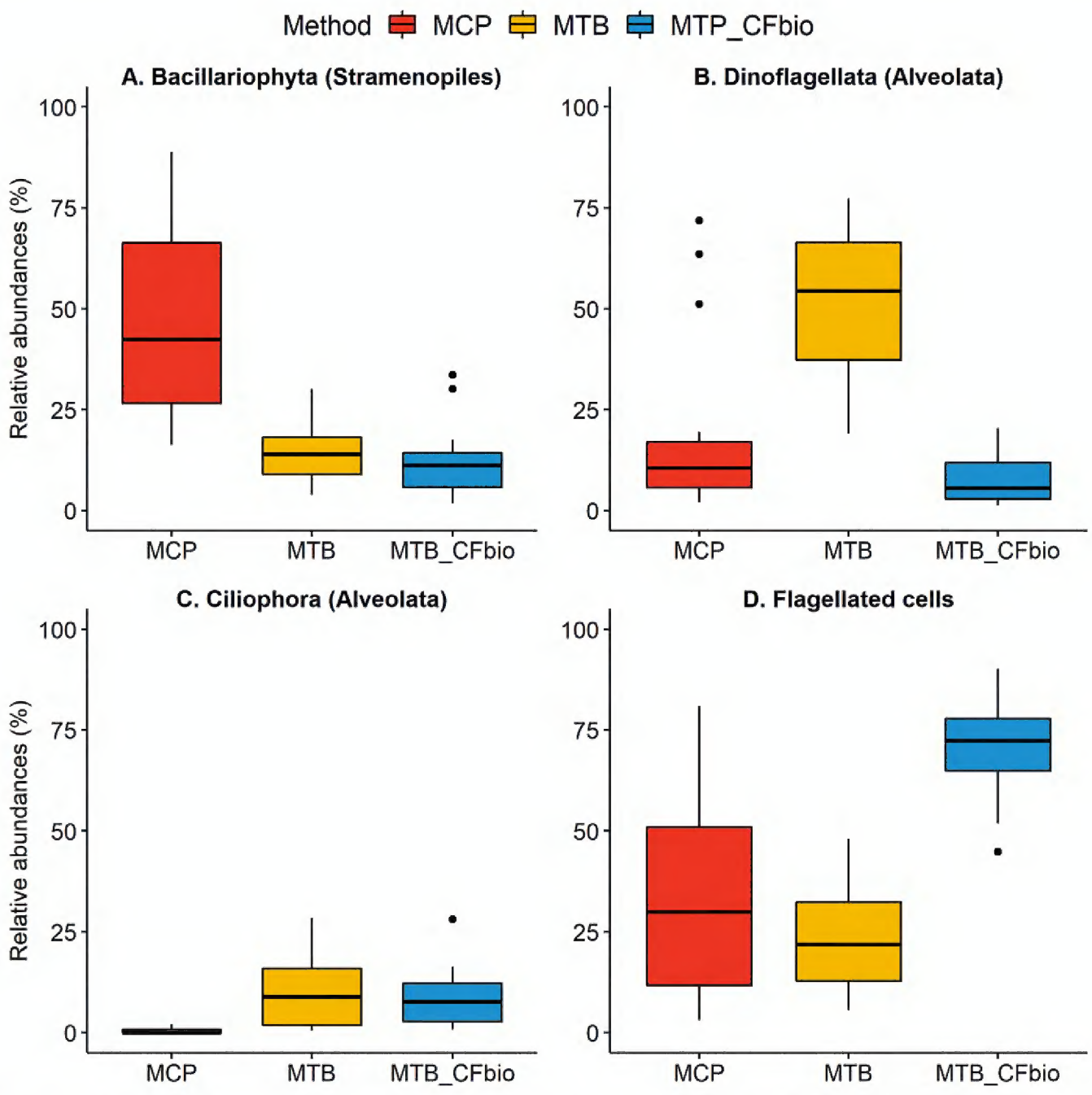


**Figure 3.** Boxplots summarizing the comparison of the 15 marine water samples biomass proportions from inverted microscopy (MCP), with DNA metabarcoding number of reads (MTB) and the corrected C-estimation values from metabarcoding reads (MTB_CFbio) using median cellular C-content values per taxum from microscopy dataset. (**A**) Bacillariophyta, (**B**) Dinoflagellata, (**C**) Ciliophora, and (**D**) flagellated cells. The vertical line inside the box plots represents the median, the top and the bottom hinges correspond to the interquartile range, and the whiskers show the minimum and maximum non-outlier values.

for Dinoflagellata. Bacillariophyta and Ciliophora displayed similar values both lower than 5 GCN:pg C.

The comparison of the biomass estimates (in relative abundances) from microscopy together with metabarcoding values and the average effect of the correction approach (MTB-CFbio) are displayed in Fig. 3. Sample per sample comparison for the entire sample set is shown in Suppl. material 4: Fig. S3. Whereas the cell relative abundances were dominated by the group flagellated cells (Fig. 3D), the C-biomass in microscopy was generally dominated by diatoms (Bacillariophyta, avg. = 46.8 ± 24%). They were followed by flagellated cells (ranged from 3 – 80.9%; avg. = 33 ± 25.2%) led by the bloom forming species *Phaeocystis globosa*, outweighing diatoms occasionally in specific samples (Samples 1,10,11,12). Dinoflagellata biomass outweighed the other groups in 3 out of 15 samples, *Noctiluca scintillans* being the main constituent of these samples (i.e.: Samples 2, 5; max. = 71.9%) with an average biomass of 19.7 ± 22.8%. Ciliate biomass estimates throughout all the samples did not attain more than 2% and was on average 0.42 ± 0.68%.

Equation (iii) was applied using CC ratios to estimate C-biomass proportions from metabarcoding reads (MTB_CFbio), which included median cellular C-content per taxum values from the microscopy dataset. Results showed that Bacillariophyta relative biomass corrections were underestimated from 42.3 ± 23.9% to 11.1 ± 9.2% (median) when compared to microscopy (Fig. 3A). Dinoflagellata estimated abundances were enhanced (Fig. 3B) from 10.4 ± 22.8% in metabarcoding to 5.6 ± 6.0% in MTB-CFbio, however, it was found that samples characterized by monospecific blooms (i.e: Samples 2, 13 and 15) were not adequately corrected (Suppl. material 4: Fig. S3). MTB-CFbio estimations of Ciliophora were similar to metabarcoding biomass proportions (MTB = 10.0 ± 8.9%; microscopy = 8.4 ± 7.6%). Both metabarcoding and MTB-CFbio strongly overestimated Ciliophora biomass compared to microscopy (Fig. 3C). Similarly, flagellated cells overall biomass was strongly overestimated by the application of the correction factor (Fig. 3D), from 29.9 ± 25.2% to 72.3 ± 12.6%.

The two beta regression analyses were performed to assess the effect of the correction factor to improve the C-biomass relative abundances. The first one compared microscopy and MTB. The second model compared MTB-CFbio using microscopy taxa specific median C-biomass dataset values (Suppl. material 3). Distributions of the two models were significant ($p < 0.01$) and detailed results are shown in Suppl. material 4: Table S4. When comparing the results of the beta regression analyses performed for microscopy proportions (displayed in Suppl. material 4: Fig. S3), and MTB-CFbio, MTB-CFbio resulted in the best fit of the beta regression model with the highest $R^2$ and Phi precision parameter ($R^2 = 0.70$ and Phi ($\varphi$) = 8.8), against the first model ($R^2$ = 059 and Phi ($\varphi$) = 7.8). The models significantly explained the variance in a different way for each phylum; this could be observed as well when examining the relative abundance estimates for each group in Suppl. material 4: Fig. S4.

# Discussion

In metabarcoding, previous attempts to control biases have primarily focused on correcting single technical biases such as fixation of the sample, DNA-extraction, group-specific primer choice, or PCR (Zarzoso-Lacoste et al. 2013; Van der Loos and Nijland 2021). As regards the correction factors implemented to improve metabarcoding quantification, various approaches have been developed targeting both micro- and macro- organisms. Thomas et al. (2016) tested the feasibility of a CF using control materials of target fish that accounted for multiple sources of bias simultaneously, and Vivien et al. (2016) exploited the consistent variations across samples between species counts sequence abundances to create a CF in aquatic oligochaetes. For prokaryotes, Angly et al. (2014) used phylogenetic differences from bacteria and archaea to develop a bioinformatic tool allowing a rapid correction of GCN in microbial surveys, resulting in improved estimates of abundances.

However, to the authors' knowledge, the only CF application in a plankton metabarcoding (freshwater) survey was carried out by Vasselon et al. (2018), and it was focused only on diatoms (Bacillariophyta). In that study, the diatom GCN and biovolume correlation was used to create a CF, which, once applied to environmental samples, served to improve the diatom quantification. Our study addressed the biological bias of 18S rRNA GCNs by accounting the taxa-specific differences in GCN. The correction factors were applied in three major unicellular eukaryotic plankton groups and for an overarching flagellate group over a 15 sample set of a marine environmental survey. However, the sampling design used to test the CF comprised a time-series, and therefore, the transferability of the proposed CF to other geographical areas is still limited. This is because the development of the CF is constrained to one environmental context and its correspondent community composition.

In addition, as previously proposed by various authors (Stern et al. 2018; Saad et al. 2020; Yarimizu et al. 2021), we contributed to the start of the first protistan 18S rRNA GCN database, generating copy number profiles for marine protists.

## Effectiveness of GCN-CF

### *Cell number*

Semi quantification by metabarcoding across groups has proven to be problematic in eukaryotes larger than 2 µm (Santoferrara et al. 2019) and good correlations between the proportion of reads and absolute species abundances in published literature are rare (Aylagas et al. 2018). To date, unfortunately, there is still little consensus on the combination of the methods for both identifying and measuring abundance of microbial populations.

Our results regarding the contrast of the two methods — microscopy (cells) and metabarcoding (sequencing

reads) — over the set of fifteen environmental samples from the Belgian Coastal Zone (Fig. 2) displayed different types of mismatches in the four groups studied: Bacillariophyta proportions between microscopy and metabarcoding were not significantly different (Fig. 2A). Conversely, flagellated cells were underrepresented in metabarcoding against microscopical observations (Fig. 2D). In the case of Alveolates (Ciliophora and Dinoflagellata), a strong overestimation of the relative abundances was observed in our environmental samples when sequence-based methodologies were used (Fig. 2B,C), but this phenomenon is well observed in the field (Elferink et al. 2017, 2020; Bruhn et al. 2021; Santi et al. 2021).

When the CF was applied, cell relative abundances were strongly improved in both alveolate groups and flagellated cells (Fig. 2). In fact, the hypothesis that one cell with many gene copies gives many reads during sequencing is addressed by applying the correction factor. However, for Bacillariophyta the application CF worsened the estimations, underestimating them. The factors explaining both microscopy and metabarcoding similarities and the CF effect worsening in Bacillariophyta may lie beyond the differences in GCNs (Fig. 1). In fact, Bacillariophyta biovolume range is the largest one of the four taxonomic groups studied (from $10^1$ to $10^9$ $\mu m^3$) (Vasselon et al. 2018), 100 times greater than for dinoflagellates (Harrison et al. 2015). Given the fact that the 18S rRNA GCN in diatoms has been significantly correlated to biovolume (Godhe et al. 2008), this huge natural variability makes the development and application of a CF challenging for Bacillariophyta. Thus, it might be necessary to apply correction factors at lower taxonomic levels in groups such as diatoms with an important morphologic diversity (i.e.: Vasselon et al. 2018).

This variability was not so obvious in the other groups studied. In accordance with the GCN database results (Fig. 1), alveolates dominated the metabarcoding dataset due to higher rRNA copy numbers than other groups. This has been already shown in previous studies (Massana et al. 2015; Käse et al. 2020; Lapeyra Martin et al. 2022). Flagellated cells underrepresentation in metabarcoding dataset is best explained by the low GCN per cell (Table 1, Fig.1). Dinoflagellata discrepancies among morphological and molecular relative abundances might be also affected by the lack of resolution of microscopical approaches to identify < 20 µm small dinoflagellates (Käse et al. 2020). Especially for ciliates, the observed discrepancy between microscopy and metabarcoding might be produced by the preservation method used.

### *C-Biomass*

There is a realization of the fact that number of reads is not generally well suited to determine absolute abundances (Medinger et al. 2010; Mäki et al. 2017). However, a meta-analysis of 22 metabarcoding studies targeting a wide variety of organisms found a weak quantitative relationship between carbon biomass and number of sequence reads produced (Lamb et al. 2019). The GCN database results displayed in Fig. 1B showed a strong positive correlation between GCN and cellular C-content for the four defined groups. This is consistent with Godhe et al. (2008) and Zhu et al. (2005) studies, that showed a significant correlation of cell length and biovolume in some marine protist taxa (Dinoflagellata and Bacillariophyta).

As regards the entire marine plankton community, Santi et al. (2021) performed a comparison between the output of microscopy analysis (in relative biomass abundances) and DNA metabarcoding, by using the same grouping of organisms as the one used in this study. Their results revealed that alveolates (Dinoflagellata and Ciliophora) displayed differences in proportions between the two methods tested whereas Bacillariophyta results did not vary significantly. The group flagellated cells showed even higher inconsistency between the two methodologies. In contrast, our results in Fig. 3 (and Suppl. material 4: Fig. S3) showed significant differences in all studied groups but flagellated cells.

The application of the CF to estimate biomass using GCN:C-content ratios, which included GCN and median C-content per taxum from the microscopy dataset, only improved the biomass estimations in two out of the four groups studied: Dinoflagellata and Ciliophora. This might be happening due to dinoflagellates that showed the highest GCN:C-content ratio and therefore the application of the correction factor resulted in the biggest impact correcting the relative abundance estimates.

The main weaknesses behind the use of GCN:C-biomass ratios for estimating biomass relative abundances from metabarcoding reads presented in this study are certainly attributed to the limitation of the number of species representing each of the plankton groups described. Particularly in the case of Ciliophora. Since there was a single ciliate species identified on our microscopy dataset, we were aware of the misrepresentation of the C-content values for the taxum Ciliophora. However, we decided to include this group in our biomass estimation analysis (application of equations ii and iii) in order to be able to perform the generalized linear modeling based on beta distribution for the entire community and assess the general effect of the correction factor.

In addition, there are more sources of bias related to the method that should be considered. For example, the estimations of C-content proportions in microscopy might be partially caused by errors in cell size measurements (and use of mean species-specific values), which are cubed to give volumes, and the effect of fixatives, which may cause shrinking or swelling of cells (Godhe et al. 2008). Additionally, there exists a difference between small cells of diatoms and dinoflagellates, which have a ~60% higher carbon density (carbon per unit cell volume) than large cells (Harrison et al. 2015). These reasons add up inaccuracies that go far beyond the 18S rRNA GCN bias in metabarcoding, and constrain the use of the correction factor to estimate relative biomass abundances.

### Limitations and future directions

DNA metabarcoding still has practical limitations (inherent to the techniques) that our GCN correction factor did not address. Some technical biases are more important than others, and amplification by PCR together with primer biases are one of the largest sources (Nichols et al. 2018). Van der Loos and Nijland (2021) suggested that PCR-free methods such as direct sequencing of rRNA genes without amplification is the obvious solution. Nonetheless, PCR dependence on metabarcoding is not likely to disappear shortly.

Two major limitations were encountered as regards the GCN database and the application of the correction factors to estimate cell and C-biomass relative abundances from metabarcoding reads. First, due to the fact that few marine protists' 18S rRNA GCNs have been assessed (Yarimizu et al. 2021) (see Suppl. material 1 and supplementary figure W3 in de Vargas et al. (2015)), we found ourselves with the necessity to use broad taxonomic phyla or plankton groups (i.e. Bacillariophyta and flagellated cells). Along with this first limitation, it must be highlighted that high variances can be found among organisms of same phyla. Nevertheless, in order not to let the outliers control the mean values, medians were used among the available values. Intra-genomic polymorphisms were used as well, which is well documented in protists (Gong et al. 2013; Pawlowski et al. 2007). Apart from its possible negative effects for species identification, it might play a role biasing the proposed CF and reducing its effectiveness.

Secondly, the GCN values used for the application of equations (i), (ii) and (iii) (see Table 1) were strictly associated with the specimens found in our GCN database, which might not necessarily coincide with a high taxonomic level resolution with the species found in our sampling area (BCZ). Accordingly, the representativeness of the defined plankton groups based on a limited number of species for each of the plankton group is constraining, likewise the replicability of the approach. Does this mean that many more gene copy number measurements of planktonic eukaryotes will enable the creation of an ideal correction factor? Fortunately, there is a considerable improvement in the measurement of 18S rRNA GCN thanks to the emergence of bioinformatic approaches that use whole genome next generation sequencing such as used by Gong and Marchetti (2019) or Sharma et al. (2021). Indeed, this tool presents promising advances — fast and cost-effective — to overcome the organismal limitations of the 18S rRNA GCN dataset. On this basis, the presented approach may become more and more robust with the unceasing increment of marine protists' GCN assessment. The GCN per cell information provided in Suppl. material 1 proceed from different methodological approaches, and confidence values were rarely reported in these studies. We believe that for the future GCN quantification studies on protists it should be considered adding exactitude parameters since some values might be more precise than others. Apart from this, it would also be interesting to consider comparing the different technical approaches and/or standardizing robust methods for future GCN quantification studies on protists. Nevertheless, other biological biases must also be pointed out due to the possible effect on the efficiency of the CF. For instance, GCN per cell will change according to the growth phase and physiological status of the cell. Indeed, Gonzalez-de-Salceda and Garcia-Pichel (2021) found that the number of 18S RNA genes per cell follows an allometric power law of cell volume with an exponent of 2/3. In addition, information regarding ribosome number variation would help to anticipate potential biases that can occur in environmental 18S rRNA metabarcoding dataset from marine environments.

Traditionally, trend analyses and dynamics of autotrophic plankton biomass have been often based on chlorophyll-*a* (Chl-*a*) pigment concentration by fluorimetry (Strickland, Parsons, 2002). To date, combination methods such as cell counting or quantitative PCR remain the only means to estimate absolute abundances (Weber and Pawlowski 2013; Canesi and Rynearson 2016; Vasselon et al. 2018; Santoferrara et al. 2020). The GCN-CF does not calculate the absolute cell numbers or C-mass per water volume, but provides relative abundances within the community. The standardization of a protocol that combines corrected metabarcoding relative abundances and Chl-*a* values might contribute to move forward towards a qualitative and quantitative monitoring of marine protistan plankton, using metabarcoding as core methodology. This way, the approach presented in this work enables us to bridge the gap between a gold standard technique in plankton research (metabarcoding) and traditional phytoplankton biomass assessment methods (Chl-*a*).

At the present time, community relative abundances are useful and reliable in the context of ecological interpretations (Piwosz et al. 2020). Even so, we believe that acknowledging the 18S rRNA GCN differences among taxa (the highest taxonomic resolution, the better) should become necessary in plankton metabarcoding over the next few years, since the application of a robust gene copy number CF will result in more accurate representations of the eukaryotic community structure. Even if clade specific values for corrections might be used as well, we must be aware that this presents its own problems due to high variances between GNC among groups (i.e.: ciliates).

## Conclusion

In the current study, we present a promising mean to measure more accurately the relative abundances of the defined marine protistan plankton groups. Our findings highlighted the need to account for these taxonomic differences in the 18S rRNA gene copy number in marine eukaryotic community studies and we proved that these might largely impact the estimates of relative abundances. We believe that the major disproportions given by biological biases in DNA-barcoding in plankton surveys may be strongly reduced using a gene copy number cor-

rection factor that can partially be explained by the differences among plankton groups.

However, the development of GCN-CF can be challenging depending on the taxum/specimen studied, as it requires finding a clear relationship between DNA reads and taxum/specimen proportions. This might be hardly possible due to the accumulation of quantification biases (e.g. cell density, cell biomass, genome size, gene copy number, intra-genomic polymorphisms) in certain marine plankton group and taxa.

The applied CF did not account for technical biases, but it greatly improved the relationship between DNA sequence read abundances and cell percentages by the traditional morphological microscopy for three out of the four groups studied, helping to normalize severe disproportions. The use of the simple and time-cost effective method presented could open a window in the meta-omics era where DNA-barcoding becomes a predominant technique to assess the major taxonomic groups both qualitatively and quantitatively. Since we are persuaded that the one-to-one relationship between 18 rRNA amplicon reads and cells/C-biomass is no longer acceptable to depict protistan plankton communities, the here presented approach not only improves the quantification approach in plankton metabarcoding, but beyond that, it opens up many new opportunities and challenges: data from eDNA metabarcoding could be easily compared and combined with the output of other approaches such as mathematical modeling of the lower trophic levels in aquatic ecosystem (often C-biomass based), which is an aspiration for many biologists.

Further investigation is needed, not only as regards the development of the eukaryotic 18S rRNA gene copy number database (which may lead to the refinement of the proposed method) but also regarding the sample set used to apply the CF, which should aim for a wider temporal and geographical scope. Accounting for these factors will help to develop correction factors to improve estimates of community abundances.

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Author contributions

JLM contributed to data acquisition and analysis, data interpretation, preparation of figures and drafting of the manuscript. NG contributed to conceptualization, resources, supervision, project administration, writing (review and editing) and funding acquisition. IS contributed to conceptualization, preparation of figures, data analysis and drafting of the manuscript. UJ and VP contributed to resources, supervision, writing (review and editing). All authors reviewed the manuscript.

## Funding

This research was supported by the MixITiN (www.mixotroph.org) project, which received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions (MSCA) grant agreement No 766327. JLM was granted with MSCA funded ITN-ETN MixITiN Early Stage Researchers (ESRs) support; JLM, NG received financial support from the Fonds David et Alice Van Buuren. UJ was financially and logistically supported through the POF IV, topic 6 and subtopic 2 research program of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research.

## Acknowledgements

We recall with affection and gratitude the dedicated technical assistance received from the crew of RV Simon Stevin and Heincke. We would like to express special thanks to the staff of the VLIZ (Flanders Marine Institute, the Hellenic Centre for Marine Research (HCMR) and Nancy Kühne, the eco-evolutionary genomics group, and AWI (Alfred Wegener Institute) for their valuable support provided.

## References

Abad D, Albaina A, Aguirre M, Laza-Martínez A, Uriarte I, Iriarte A, Villate F, Estonba A (2016) Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. Marine Biology 163(7): 149. https://doi.org/10.1007/s00227-016-2920-0

Angly F, Dennis P, Skarshewski A, Vanwonterghem I, Philip H, Tyson G (2014) Copyrighter: A rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2(1): 11. https://doi.org/10.1186/2049-2618-2-11

Armeli Minicante S, Piredda R, Quero GM, Finotto S, Bernardi Aubry F, Bastianini M, Pugnetti A, Zingone A (2019) Habitat Heterogeneity and Connectivity: Effects on the Planktonic Protist Community Structure at Two Adjacent Coastal Sites (the Lagoon and the Gulf of Venice, Northern Adriatic Sea, Italy) Revealed by Metabarcoding. Frontiers in Microbiology 10: 2736. https://doi.org/10.3389/fmicb.2019.02736

Aylagas E, Rodriguez-Ezpeleta N, Borja A (2018) Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. Ecological Indicators 95: 194–202. https://doi.org/10.1016/j.ecolind.2018.07.044

Bruhn CS, Wohlrab S, Krock B, Lundholm N, John U (2021) Seasonal plankton succession is in accordance with phycotoxin occurrence in Disko Bay, West Greenland. Harmful Algae 103: 101978. https://doi.org/10.1016/j.hal.2021.101978

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13(7): 581–583. https://doi.org/10.1038/nmeth.3869

Canesi KL, Rynearson TA (2016) Temporal variation of Skeletonema community composition from a long-term time series in Narragansett Bay identified using high-throughput DNA sequencing. Marine Ecology Progress Series 556: 1–16. https://doi.org/10.3354/meps11843

Chain FJJ, Brown EA, MacIsaac HJ, Cristescu ME (2016) Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. Diversity & Distributions 22(5): 493–504. https://doi.org/10.1111/ddi.12427

Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, Moline MA (2008) Correlated Evolution of Genome Size and Cell Volume in Diatoms (bacillariophyceae)1. Journal of Phycology 44(1): 124–131. https://doi.org/10.1111/j.1529-8817.2007.00452.x

Cribari-Neto F, Zeileis A (2010) Beta regression in R. Journal of Statistical Software 34(2): 1–24. https://doi.org/10.18637/jss.v034.i02

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E, Boss E, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sullivan MB, Velayoudon D (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348(6237): 1261605. https://doi.org/10.1126/science.1261605

Ebenezer V, Medlin LK, Ki J-S (2012) Molecular Detection, Quantification, and Diversity Evaluation of Microalgae. Marine Biotechnology 14(2): 129–142. https://doi.org/10.1007/s10126-011-9427-y

Edler L, Elbrächter M (2010) The Utermöhl method for quantitative phytoplankton analysis. Microscopic and molecular methods for quantitative phytoplankton analysis 110: 13–20.

Elbrecht V, Leese F (2015) Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLoS ONE 10(7): e0130324. https://doi.org/10.1371/journal.pone.0130324

Elferink S, Neuhaus S, Wohlrab S, Toebe K, Voß D, Gottschling M, Lundholm N, Krock B, Koch BP, Zielinski O, Cembella A, John U (2017) Molecular diversity patterns among various phytoplankton size-fractions in West Greenland in late summer. Deep-sea Research. Part I, Oceanographic Research Papers 121: 54–69. https://doi.org/10.1016/j.dsr.2016.11.002

Elferink S, Wohlrab S, Neuhaus S, Cembella A, Harms L, John U (2020) Comparative Metabarcoding and Metatranscriptomic Analysis of Microeukaryotes Within Coastal Surface Waters of West Greenland and Northwest Iceland. Frontiers in Marine Science 7. https://www.frontiersin.org/article/10.3389/fmars.2020.00439

Ershova EA, Wangensteen OS, Descoteaux R, Barth-Jensen C, Præbel K (2021) Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. ICES Journal of Marine Science 78(9): 3342–3355. https://doi.org/10.1093/icesjms/fsab171

Galluzzi L, Penna A (2013) Quantitative PCR for detection and enumeration of phyto-plankton. Microscopic and molecular methods for quantitative phytoplankton analysis 95.

Godhe A, Asplund ME, Härnström K, Saravanan V, Tyagi A, Karunasagar I (2008) Quantification of Diatom and Dinoflagellate Biomasses in Coastal Marine Seawater Samples by Real-Time PCR. Applied and Environmental Microbiology 74(23): 7174–7182. https://doi.org/10.1128/AEM.01298-08

Gong W, Marchetti A (2019) Estimation of 18S Gene Copy Number in Marine Eukaryotic Plankton Using a Next-Generation Sequencing Approach. Frontiers in Marine Science 6: 219. https://doi.org/10.3389/fmars.2019.00219

Gong J, Dong J, Liu X, Massana R (2013) Extremely High Copy Numbers and Polymorphisms of the rDNA Operon Estimated from Single Cell Analysis of Oligotrich and Peritrich Ciliates. Protist 164(3): 369–379. https://doi.org/10.1016/j.protis.2012.11.006

Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A (2012) Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities. PLoS ONE 7(1): e29973. https://doi.org/10.1371/journal.pone.0029973

Gonzalez-de-Salceda L, Garcia-Pichel F (2021) The allometry of cellular DNA and ribosomal gene content among microbes and its use for the assessment of microbiome community structure. Microbiome 9(1): 173. https://doi.org/10.1186/s40168-021-01111-z

Gran-Stadniczeñko S, Egge E, Hostyeva V, Logares R, Eikrem W, Edvardsen B (2019) Protist Diversity and Seasonal Dynamics in Skagerrak Plankton Communities as Revealed by Metabarcoding and Microscopy. The Journal of Eukaryotic Microbiology 66(3): 494–513. https://doi.org/10.1111/jeu.12700

Groendahl S, Kahlert M, Fink P (2017) The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. PLoS ONE 12(2): e0172808. https://doi.org/10.1371/journal.pone.0172808

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R (2013) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41(D1): D597–D604. https://doi.org/10.1093/nar/gks1160

Gypens N, Lacroix G, Lancelot C (2007) Causes of variability in diatom and *Phaeocystis* blooms in Belgian coastal waters between 1989 and 2003: A model study. Journal of Sea Research 57(1): 19–35. https://doi.org/10.1016/j.seares.2006.07.004

Harrison PJ, Zingone A, Mickelson MJ, Lehtinen S, Ramaiah N, Kraberg AC, Sun J, McQuatters-Gollop A, Jakobsen HH (2015) Cell volumes of marine phytoplankton from globally distributed coastal data sets. Estuarine, Coastal and Shelf Science 162: 130–142. https://doi.org/10.1016/j.ecss.2015.05.026

Illumina I (2013) 16S Metagenomic sequencing library preparation. Preparing 16S ribosomal RNA gene amplicons for the illumina MiSeq system 1: 28.

Käse L, Kraberg AC, Metfies K, Neuhaus S, Sprong PAA, Fuchs BM, Boersma M, Wiltshire KH (2020) Rapid succession drives spring community dynamics of small protists at Helgoland Roads, North Sea. Journal of Plankton Research 42(3): 305–319. https://doi.org/10.1093/plankt/fbaa017

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Computational Biology 8(10): e1002743. https://doi.org/10.1371/journal.pcbi.1002743

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7(1): 17668. https://doi.org/10.1038/s41598-017-17333-x

LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW (2005) Symbiodinium (pyrrhophyta) Genome Sizes (dna Content) Are Smallest Among Dinoflagellates1. Journal of Phycology 41(4): 880–886. https://doi.org/10.1111/j.0022-3646.2005.04231.x

Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2019) How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology 28(2): 420–430. https://doi.org/10.1111/mec.14920

Lapeyra Martin J, John U, Royer C, Gypens N (2022) Fantastic Beasts: Unfolding Mixoplankton Temporal Variability in the Belgian Coastal Zone Through DNA-Metabarcoding. Frontiers in Marine Science 9: 786787. https://doi.org/10.3389/fmars.2022.786787

Latz MAC, Grujcic V, Brugel S, Lycken J, John U, Karlson B, Andersson A, Andersson AF (2022) Short- and long-read metabarcoding of the eukaryotic rRNA operon: evaluation of primers and comparison to shotgun metagenomics sequencing. Molecular Ecology Resources: 1755–0998.13623. https://doi.org/10.1111/1755-0998.13623

Lee ZM-P, Bussema C III, Schmidt TM (2009) rrnDB: Documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Research 37(Database): D489–D493. https://doi.org/10.1093/nar/gkn689

Mäki A, Salmi P, Mikkonen A, Kremp A, Tiirola M (2017) Sample Preservation, DNA or RNA Extraction and Data Analysis for High-Throughput Phytoplankton Community Sequencing. Frontiers in Microbiology 8: 1848. https://doi.org/10.3389/fmicb.2017.01848

Manoylov KM (2014) Taxonomic identification of algae (morphological and molecular): Species concepts, methodologies, and their implications for ecological bioassessment. Journal of Phycology 50(3): 409–424. https://doi.org/10.1111/jpy.12183

Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, Chambouvet A, Christen R, Claverie J-M, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Forn I, Forster D, Guillou L, Jaillon O, Kooistra WHCF, Logares R, Mahé F, Not F, Ogata H, Pawlowski J, Pernice MC, Probert I, Romac S, Richards T, Santini S, Shalchian-Tabrizi K, Siano R, Simon N, Stoeck T, Vaulot D, Zingone A, de Vargas C (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environmental Microbiology 17(10): 4035–4049. https://doi.org/10.1111/1462-2920.12955

Matesanz S, Pescador DS, Pías B, Sánchez AM, Chacón-Labella J, Illuminati A, de la Cruz M, López-Angulo J, Marí-Mena N, Vizcaíno A, Escudero A (2019) Estimating belowground plant abundance with DNA metabarcoding. Molecular Ecology Resources 19(5): 1265–1277. https://doi.org/10.1111/1755-0998.13049

Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J (2010) Diversity in a hidden world: Potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Molecular Ecology 19: 32–40. https://doi.org/10.1111/j.1365-294X.2009.04478.x

Medlin LK, Kooistra WHCF (2010) Methods to Estimate the Diversity in the Marine Photosynthetic Protist Community with Illustrations from Case Studies: A Review. Diversity 2(7): 973–1014. https://doi.org/10.3390/d2070973

Menden-Deuer S, Lessard EJ (2000) Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. Limnology and Oceanography 45(3): 569–579. https://doi.org/10.4319/lo.2000.45.3.0569

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, Shapiro B (2018) Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources 18(5): 927–939. https://doi.org/10.1111/1755-0998.12895

Nohe A, Knockaert C, Goffin A, Dewitte E, De Cauwer K, Desmit X, Vyverman W, Tyberghein L, Lagring R, Sabbe K (2018) Marine phytoplankton community composition data from the Belgian part of the North Sea, 1968–2010. Scientific Data 5(1): 180126. https://doi.org/10.1038/sdata.2018.126

Not F, del Campo J, Balagué V, de Vargas C, Massana R (2009) New Insights into the Diversity of Marine Picoeukaryotes. PLoS ONE 4(9): e7143. https://doi.org/10.1371/journal.pone.0007143

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2013) Package 'vegan.' Community ecology package, version 2: 1–295.

Olenina I, Hajdu S, Edler L, Andersson A, Wasmund N, Busch S, Göbel J, Gromisz S, Huseby S, Huttunen M, Jaanus A, Kokkonen P, Ledaine I, Niemkiewicz E (2006) Biovolumes and size-classes of phytoplankton in the Baltic Sea. HELCOM. Baltic Sea Environment Proceedings 106: 1–144.

Pawlowski J, Fahrni J, Lecroq B, Longet D, Cornelius N, Excoffier L, Cedhagen T, Gooday AJ (2007) Bipolar gene flow in deep-sea benthic foraminifera. Molecular Ecology 16(19): 4089–4096. https://doi.org/10.1111/j.1365-294X.2007.03465.x

Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, Visco J, Esling P (2016) Protist metabarcoding and environmental biomonitoring: Time for change. European Journal of Protistology 55: 12–25. https://doi.org/10.1016/j.ejop.2016.02.003

Peuto-Moreau M (1991) Symbiose plastidiale et mixotrophie des cilies planctoniques marins (ciliophora oligotrichina). These de doctorat, Nice. https://www.theses.fr/1991NICE4473

Piredda R, Tomasino MP, D'Erchia AM, Manzari C, Pesole G, Montresor M, Kooistra WHCF, Sarno D, Zingone A (2017) Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. FEMS Microbiology Ecology 93(1): fiw200. https://doi.org/10.1093/femsec/fiw200

Piwosz K, Shabarova T, Pernthaler J, Posch T, Šimek K, Porcal P (2020) Salcher MM Bacterial and Eukaryotic Small-Subunit Amplicon Data Do Not Provide a Quantitative Picture of Microbial Communities, but They Are Reliable in the Context of Ecological Interpretations. MSphere 5: e00052–e20. https://doi.org/10.1128/mSphere.00052-20

Putt M, Stoecker DK (1989) An experimentally determined carbon : Volume ratio for marine "oligotrichous" ciliates from estuarine and coastal waters. Limnology and Oceanography 34(6): 1097–1103. https://doi.org/10.4319/lo.1989.34.6.1097

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A, Young J, Aguilar M, Claverie J-M, Frickenhaus S, Gonzalez K, Herman EK, Lin Y-C, Napier J, Ogata H, Sarno AF, Shmutz J, Schroeder D, de Vargas C, Verret F, von Dassow P, Valentin K, Van de Peer Y, Wheeler G, Dacks JB, Delwiche CF, Dyhrman ST, Glöckner G, John U, Richards T, Worden AZ, Zhang X, Grigoriev IV (2013) Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499(7457): 209–213. https://doi.org/10.1038/nature12221

Rousseau V, Mathot S, Lancelot C (1990) Calculating carbon biomass of *Phaeocystis* sp. from microscopic observations. Marine Biology 107(2): 305–314. https://doi.org/10.1007/BF01319830

Saad OS, Lin X, Ng TY, Li L, Ang P, Lin S (2020) Genome Size, rDNA Copy, and qPCR Assays for Symbiodiniaceae. Frontiers in Microbiology 11. https://doi.org/10.3389/fmicb.2020.00847

Santi I, Kasapidis P, Karakassis I, Pitta P (2021) A Comparison of DNA Metabarcoding and Microscopy Methodologies for the Study of

Aquatic Microbial Eukaryotes. Diversity 13(5): 180. https://doi.org/10.3390/d13050180

Santoferrara LF (2019) Current practice in plankton metabarcoding: Optimization and error management. Journal of Plankton Research 41(5): 571–582. https://doi.org/10.1093/plankt/fbz041

Santoferrara LF, Burki F, Filker S, Logares R, Dunthorn M, McManus GB (2020) Perspectives from Ten Years of Protist Studies by High-Throughput Metabarcoding. The Journal of Eukaryotic Microbiology 67(5): 612–622. https://doi.org/10.1111/jeu.12813

Sharma D, Denmat SH-L, Matzke NJ, Hannan K, Hannan RD, O'Sullivan JM, Ganley ARD (2021) A new method for determining ribosomal DNA copy number shows differences between Saccharomyces cerevisiae populations. bioRxiv, 1–43. https://doi.org/10.1101/2021.01.21.427686

Smithson M, Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods 11(1): 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Stern R, Kraberg A, Bresnan E, Kooistra WHCF, Lovejoy C, Montresor M, Morán XAG, Not F, Salas R, Siano R, Vaulot D, Amaral-Zettler L, Zingone A, Metfies K (2018) Molecular analyses of protists in long-term observation programmes—Current status and future perspectives. Journal of Plankton Research 40(5): 519–536. https://doi.org/10.1093/plankt/fby035

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular Ecology 19: 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Strickland JDH, Parsons TR (2002) A practical handbook of seawater analysis. https://publications.gc.ca/site/eng/480760/publication.html

R Team (2018) R: A language and environment for statistical computing.

Thomas AC, Deagle BE, Eveson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources 16(3): 714–726. https://doi.org/10.1111/1755-0998.12490

Toebe K, Joshi AR, Messtorff P, Tillmann U, Cembella A, John U (2013) Molecular discrimination of taxa within the dinoflagellate genus Azadinium, the source of azaspiracid toxins. Journal of Plankton Research 35(1): 225–230. https://doi.org/10.1093/plankt/fbs077

Van der Loos LM, Nijland R (2021) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology 30(13): 3270–3288. https://doi.org/10.1111/mec.15592

Vasselon V, Bouchez A, Rimet F, Jacquet S, Trobajo R, Corniquel M, Tapolczai K, Domaizon I (2018) Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. Methods in Ecology and Evolution 9(4): 1060–1069. https://doi.org/10.1111/2041-210X.12960

Vivien R, Lejzerowicz F, Pawlowski J (2016) Next-Generation Sequencing of Aquatic Oligochaetes: Comparison of Experimental Communities. PLoS ONE 11(2): e0148644. https://doi.org/10.1371/journal.pone.0148644

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73(16): 5261–5267. https://doi.org/10.1128/AEM.00062-07

Weber AA-T, Pawlowski J (2013) Can Abundance of Protists Be Inferred from Sequence Data: A Case Study of Foraminifera. PLoS ONE 8(2): e56739. https://doi.org/10.1371/journal.pone.0056739

Weisse T, Anderson R, Arndt H, Calbet A, Hansen PJ, Montagnes DJS (2016) Functional ecology of aquatic phagotrophic protists – Concepts, limitations, and perspectives. European Journal of Protistology 55: 50–74. https://doi.org/10.1016/j.ejop.2016.03.003

Wickham H (2011) ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3(2): 180–185. https://doi.org/10.1002/wics.147

Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews 21: 213–229. https://doi.org/10.1111/j.1574-6976.1997.tb00351.x

Yarimizu K, Sildever S, Hamamoto Y, Tazawa S, Oikawa H, Yamaguchi H, Basti L, Mardones JI, Paredes-Mella J, Nagai S (2021) Development of an absolute quantification method for ribosomal RNA gene copy numbers per eukaryotic single cell by digital PCR. Harmful Algae 103: 102008. https://doi.org/10.1016/j.hal.2021.102008

Zarzoso-Lacoste D, Corse E, Vidal E (2013) Improving PCR detection of prey in molecular diet studies: Importance of group-specific primer set selection and extraction protocol performances. Molecular Ecology Resources 13(1): 117–127. https://doi.org/10.1111/1755-0998.12029

Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiology Ecology 52(1): 79–92. https://doi.org/10.1016/j.femsec.2004.10.006

**Supplementary material 1**
**Supplementary Data 1**

Author: Jon Lapeyra Martin, Ioulia Santi, Paraskevi Pitta, Uwe John, Nathalie Gypens

Data type: morphological, genomic

Explanation note: Gene copy number dataset containing 18S rRNA copy number, biovolume and C-content of species.



Link: https://doi.org/10.3897/mbmg.6.85794.suppl1

**Supplementary material 2**
**Supplementary Data 2**

Author: Jon Lapeyra Martin, Ioulia Santi, Paraskevi Pitta, Uwe John, Nathalie Gypens

Data type: genomic, occurences

Explanation note: ASV counts per sample of the 15 sample set (time-series) from the Belgial Coastal Zone.



Link: https://doi.org/10.3897/mbmg.6.85794.suppl2

**Supplementary material 3**
**Supplementary Data 3**
Author: Jon Lapeyra Martin, Ioulia Santi, Paraskevi Pitta, Uwe John, Nathalie Gypens
Data type: occurences, morphological
Explanation note: Summary of the species identified in microscopy dataset, biovolumes and biomass estimations.

Link: https://doi.org/10.3897/mbmg.6.85794.suppl3

**Supplementary material 4**
**Tables S1–S4, Figures S1–S4**
Author: Jon Lapeyra Martin, Ioulia Santi, Paraskevi Pitta, Uwe John, Nathalie Gypens
Data type: Figures, tables
Explanation note: Supplementary tables and figures.

Link: https://doi.org/10.3897/mbmg.6.85794.suppl4